

MOSCARD - Multimodal Opportunistic Screening for Cardiovascular Adverse events with Causal Reasoning and De-confounding

Jialu Pi¹, Juan Maria Farina², Rimita Lahiri³, Jiwoong Jeong¹, Archana Gurudu³, Hyung-Bok Park⁵, Chieh-Ju Chao⁴, Chadi Ayoub², Reza Arsanjani², and Imon Banerjee^{1,3}

¹ SCAI, Arizona State University, Tempe, AZ, USA

² Department of Cardiovascular Medicine, Mayo Clinic, Scottsdale, AZ, USA

³ Department of Radiology, Mayo Clinic, Phoenix, AZ, USA

⁴ Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA

⁵ Department of Cardiology, International St. Mary's Hospital, Catholic Kwandong University, South Korea

{jialupi, jjeong35}@asu.edu, {argurudu14, hyungbok7}@gmail.com,
{Farina.JuanMaria, Lahiri.Rimita, Chao.ChieJu, Ayoub.Chadi,
Arsanjani.Reza, Banerjee.Imon}@mayo.edu

Abstract. Major Adverse Cardiovascular Events (MACE) remain the leading cause of mortality globally, as reported in the Global Disease Burden Study 2021. Opportunistic screening leverages data collected from routine health check-ups and multimodal data can play a key role to identify at-risk individuals. Chest X-rays (CXR) provide insights into chronic conditions contributing to major adverse cardiovascular events (MACE), while 12-lead electrocardiogram (ECG) directly assesses cardiac electrical activity and structural abnormalities. Integrating CXR and ECG could offer a more comprehensive risk assessment than conventional models, which rely on clinical scores, computed tomography (CT) measurements, or biomarkers, which may be limited by sampling bias and single modality constraints. We propose a novel predictive modeling framework - MOSCARD, multimodal causal reasoning with co-attention to align two distinct modalities and simultaneously mitigate bias and confounders in opportunistic risk estimation. Primary technical contributions are - (i) multimodal alignment of CXR with ECG guidance; (ii) integration of causal reasoning; (iii) dual back-propagation graph for de-confounding. Evaluated on internal, shift data from emergency department (ED) and external MIMIC datasets, our model outperformed single modality and state-of-the-art foundational models - AUC: 0.75, 0.83, 0.71 respectively. Proposed cost-effective opportunistic screening enables early intervention, improving patient outcomes and reducing disparities⁶.

Keywords: Major adverse cardiovascular events · Multimodal learning · Causal reasoning · Predictive models.

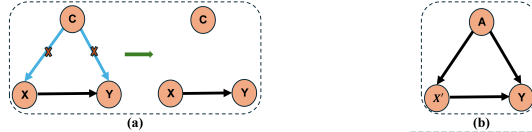
⁶ Code will be available at <https://github.com/OrchidPi/MOSCARD>

1 Introduction

Major adverse cardiovascular events (MACE) are among the leading causes of death worldwide [11]. Opportunistic screening refers to identifying individuals at risk of MACE during routine health check-ups leveraging multimodal data to provide timely intervention for high and medium risk individual. Although modern clinical practice is highly dependent on the synthesis of information from multiple heterogeneous data sources, traditional predictive models rely on clinical data [18, 16] or CT-based imaging biomarkers [6, 21, 19]. Combination of heterogeneous high-dimensional data is still a substantial technical challenge in healthcare AI due to data complexity, noise, semantic interoperability, bias, and ethical dilemmas [1]. Recent advancements with the vision-language models (e.g., MedClip) remains limited in medical domain due to constraints of the modalities (primarily text and image) and the inherent biases present [14]. Furthermore, these models often exhibit reduced performance in diverse populations due to population bias based on selective sampling [10, 4]. This bias occurs when training data lack sufficient diversity, resulting in shortcut learning and consistent inaccuracies between underrepresented groups [2, 3, 5].

While Chest X-ray (CXR) is particularly useful for detecting signs of chronic conditions that contribute to cardiac disease, 12-lead electrocardiogram (ECG) provides a direct assessment of the electrical activity of the heart, detecting arrhythmia, ischemic changes, and structural abnormalities. CXR integrated with 12-lead ECG captured during the same visit could represent a more holistic view of a patient’s cardiac health. For opportunistic screening, this combination could be especially useful in settings where resources or access to more sophisticated imaging like echocardiography, CT, magnetic resonance imaging (MRI) are limited. Existing AI models, while effective, often learn demographic biases, leading to disparities in disease prediction. A recent review on ‘fair’ AI model [9] highlights three key methodologies - fair dataset curation, unbiased representation learning, and adversarial debiasing. Achieving fair dataset by ensuring equal representation of all sensitive attributes can be challenging when data is collected from the real-world distribution. Adversarial debiasing has been shown to reduce racial bias without the accessing balanced dataset [8], however, its complexity limits its effectiveness, as it typically targets only single confounder, failing to model interdependent factors such as age and comorbidities. We propose a multimodal causal reasoning framework - **MOSCARD** to address the impact of bias and confounders for opportunistic screening for composite short (6 months) and long (5 years) term MACE. The proposed model leverages the posterior-anterior view (PA) of CXR images, and 12-lead ECGs to learn ‘fair’ representation while modeling causality to calculate the probabilistic risk score.

Our primary contributions are - (i) novel co-alignment framework to integrate the knowledge from ECG data to guide the embedding space alignment for CXR, (ii) remove the confounder effect (selection bias) for CXR and ECG data featurization through confusion loss, (iii) apply causal reasoning to model the causal relationship between comorbidities and the targeted MACE risk (e.g. Obesity $\rightarrow$ Insulin Resistance $\rightarrow$ Dyslipidemia $\rightarrow$ Atherosclerosis $\rightarrow$ CVD Risk).



MOSCARD is compared against the single modality and state-of-the-art multi-modal deep learning models [20, 15] without debiasing and causal intervention across multiple datasets (including MIMICIV [13]) with and without the domain shift.

We proposed a two step multimodal framework - MOSCARD, where in step 1 we train the single modality encoders with confusion loss to remove confounder effect (Fig. 1 a) and in step 2 we apply multimodal learning with co-attention (Fig. 1 b). For modeling causal relationships in step 2, we applied a Structural Causal Model (SCM) and their parametrization as Directed Acyclical Graphs (DAG) [17] representing the image (X), target label (Y), causal factor (A), and confounder (C). The multimodal model consists of three prediction branches: the main branch is for predicting MACE (Y), while the other two discriminative branches are predicting causal factors (Y_{ca}) and confounder labels (Y_{conf}) using features from the early transformer layers of the vision transformer encoder.

For step 1, we design the de-confounding through distinct 12 layers ViT encoders for processing CXR and ECG images (X_{CXR} & X_{ECG}). The causal relationship $X \rightarrow Y$ represents the desired effect, however the backdoor path $X \leftarrow C \rightarrow Y$ highlights the influence of the confounder C , which introduces spurious correlations and affects predictions (Figure 1 a). To mitigate this bias, we remove the confounding influence separately from each encoder by eliminating the wrong links $C \rightarrow X$ and $C \rightarrow Y$ using two separate backpropagation graphs. The first

loss function focuses on the main MACE predictions task: $\text{Loss}_{\text{Main}} = L_{\text{CE}}(y, \hat{y})$ and the second loss is dedicated to confounder task, handling gender and age predictions using cross-entropy: $\text{Loss}_{\text{Conf}} = L_{\text{CE}}(y_c, \hat{y}_c) + \alpha L_{\text{Conf}}(\hat{y}_c)$. To further mitigate confounder influence, we incorporate confusion loss $\alpha L_{\text{Conf}}(\hat{y}_c)$, encouraging the confounder predictions to approximate a uniform distribution [12]. To effectively mitigate confounders, we utilize features from the early layers of the ViT which primarily capture low-level patterns and naturally encode strong confounding signals related to factors like gender and age.

2.2 Multimodal Co-attention alignment

We utilize the pre-trained encoder weights from Step 1, freezing them in the subsequent co-attention module (Fig. 1 d). However, the primary challenge for the ECG and CXR alignment is unknown spatial relation. To preserve the primary knowledge from CXR and use ECG as a guiding modality, we adapted the co-attention mechanism in H&E slides processing with genomic factor [7]. ECG features (E) serve as the query, while CXR features (CXR) function as both the key and value in the co-attention computation: $\text{CoAttn}_{E \rightarrow CXR}(E, C) = \text{softmax}\left(\frac{Q_{CXR} K_E^\top}{\sqrt{d_k}}\right) V_E$. The attention scores are obtained by computing the similarity between the projected query ($Q = W_q^E E$) and key ($K = W_k^{CXR} CXR$), scaled by d_k for numerical stability. A softmax operation is applied to produce the co-attention matrix ($A_{\text{coattn}} \in \mathbb{R}^{N \times M}$), which assigns importance weights to different regions of CXR . These attention weights are used to compute a weighted sum of the CXR features, refining them into $\hat{CXR}$ as:

$$\text{softmax}\left(\frac{W_q^E E C X R^\top W_k^{C X R}}{\sqrt{d_k}}\right) W_v^{C X R} C X R \rightarrow A_{\text{coattn}} W_v^{C X R} C X R \rightarrow \hat{C X R}$$

, where $W_q^E, W_k^{CXR}, W_v^{CXR} \in \mathbb{R}^{d_k \times d_k}$ are trainable weights. The values are derived from the CXR features as $V = W_v^{CXR} CXR$ which ensures that the model selectively attends to the most informative regions, conditioned on ECG features.

After the co-attention mechanism, the CXR and ECG features are processed through two transformers and multi-layer perceptron (MLP) branch to perform the classification tasks. The model consists of five branches, addressing two distinct tasks: the causal task, which predicts the classification label Y_{ca} , and the main task, which predicts the classification label Y . As shown in Fig. 1 b the relationship $X \leftarrow A \rightarrow Y$ indicates that causal factors A influence both X and Y . Our goal is to allow the model to leverage information from A and improve generalization in MACE prediction. Causal learning is applied separately to both CXR and ECG features to preserve the heterogeneity of modality-specific information and their combined representations for improved classification performance. The loss function of five branches classification task is presented as: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{causal}}^{\text{CXR}} + \mathcal{L}_{\text{main}}^{\text{CXR}} + \mathcal{L}_{\text{causal}}^{\text{ECG}} + \mathcal{L}_{\text{main}}^{\text{ECG}} + \mathcal{L}_{\text{main}}^{\text{concat}}$, where $\mathcal{L}_{\text{causal}}^{\text{CXR}}$ and $\mathcal{L}_{\text{main}}^{\text{CXR}}$ denote the causal and main losses for CXR, $\mathcal{L}_{\text{causal}}^{\text{ECG}}$ and $\mathcal{L}_{\text{main}}^{\text{ECG}}$ represent the causal and main losses for ECG, and $\mathcal{L}_{\text{main}}^{\text{concat}}$ is the combined loss for concatenated ECG and CXR.

3 Results

Internal cohort. Upon approval Institutional Review Board, retrospective data for 21,872 patients were collected across geographically disparate four sites. Patients who were lost to follow-up within 6 months were excluded due to uncertainty regarding the outcomes. Subsequent MACE outcomes were compiled by manual chart-review. Ultimately, the cohort included 12,612 patients who CXR performed and had a matching 12-lead ECG within a ± 6 months time window. The dataset was randomly split at the patient level (80% training, 20% testing), yielding 79,648 and 21,108 image pairs, respectively. The dataset consists of high-risk cardiac patients, with elevated 51.77% and 55.91% MACE event rate within 6 month and one year respectively.

Shift cohort (ED). To evaluate the model’s generalization across ‘shift’ data, we collected testing data from patient with CXR for any indication performed at Emergency Department (ED). The Shift cohort consisted of 2,638 patients and ECG signals were matched with CXR yielding 2,830 image-ECG pair. Compared to training population, this cohort includes younger patients (8.34% over 80yrs) with significantly low presence of comorbidities, such as CHF (0%), CKD (0%), diabetes (0.94%) and hypertension (2%), as well as low MACE rate (25.12%) within one year.

Shift cohort (MIMIC). To benchmark, we evaluated the model on publicly available MIMICIV eICU data [13]. We randomly selected 200 patients who underwent inpatient CXR imaging for any clinical indication. Among them, we identified 175 patients with temporally matched ECG recordings. We curated the 6 month MACE outcome by parsing the clinic notes and inpatient mortality data and MACE rate was 25.71%.

Performance: We compared single modality and baseline multimodal models against our proposed confounder debiasing (Conf), causal+confounder (With Confounder), and causal only (Without Confounder) models. The baseline single modality encoders are derived from MedCLIP and are fine-tuned separately for CXR and ECG data (CXR and ECG model respectively). We compared the proposed multimodal alignment framework MOSCARD against - MedCLIP (contrastive) and ALBEF (knowledge distillation) baselines [20, 15]. The evaluation results for single and multimodality models are summarized in Table 1 and 2.

On the internal datasets, *Baseline CXR* achieves the best performance, particularly for MACE prediction at 5 years, with an AUC of 0.739 (Table 1). However, incorporating the de-confounding branch performance does not significantly improve on the internal testset, indicating that de-confounding does not enhance predictive performance on the hold-out test; however, the performance is on the par with the baseline (CXR Conf 5 years AUC 0.722). De-confounding models demonstrate optimal generalizability on external datasets (ED: 0.778 CXR Conf and MIMIC: 0.66 AUC ECG Conf), emphasizing that baseline models without de-confounding may rely on shortcut features for prediction, which are not generalizable.

For multimodal models, MOSCARD model with causality achieves the best performance on both internal and external datasets and outperformed MedCLIP

Table 1: Comparison of single modality models with and without the de-confounding (95% confidence intervals using bootstrapping). Green cells highlight the optimal performance.

Single Modality	CXR		ECG		CXR Conf		ECG Conf		
	MACE	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	
Internal	6M	0.656 [0.653, 0.659]	0.710 [0.707, 0.715]	0.594 [0.593, 0.598]	0.633 [0.625, 0.634]	0.639 [0.639, 0.646]	0.694 [0.693, 0.700]	0.607 [0.604, 0.612]	0.648 [0.646, 0.654]
	1yr	0.664 [0.660, 0.668]	0.715 [0.711, 0.717]	0.599 [0.594, 0.601]	0.640 [0.637, 0.647]	0.646 [0.640, 0.647]	0.698 [0.693, 0.701]	0.616 [0.611, 0.618]	0.656 [0.656, 0.661]
	2yr	0.665 [0.666, 0.673]	0.719 [0.713, 0.724]	0.605 [0.603, 0.611]	0.644 [0.642, 0.647]	0.651 [0.649, 0.656]	0.702 [0.699, 0.710]	0.618 [0.612, 0.620]	0.657 [0.655, 0.661]
	5yr	0.679 [0.679, 0.684]	0.735 [0.735, 0.745]	0.607 [0.602, 0.607]	0.645 [0.645, 0.652]	0.665 [0.649, 0.670]	0.722 [0.715, 0.725]	0.614 [0.611, 0.617]	0.661 [0.660, 0.669]
	External (MIMIC)	0.560 [0.546, 0.575]	0.642 [0.630, 0.658]	0.663 [0.663, 0.687]	0.699 [0.694, 0.725]	0.500 [0.547, 0.568]	0.612 [0.600, 0.627]	0.651 [0.645, 0.668]	0.660 [0.645, 0.679]
External (ED)	1yr	0.724 [0.720, 0.727]	0.768 [0.764, 0.777]	0.673 [0.669, 0.676]	0.742 [0.733, 0.749]	0.712 [0.710, 0.716]	0.778 [0.774, 0.784]	0.670 [0.666, 0.676]	0.724 [0.712, 0.726]

Table 2: Comparative model performance after multimodal data alignment with combined and individual modality (95% confidence intervals using bootstrapping). Green cells highlights the optimal performance.

Knowledge distillation Baseline (ALBEF)							
	MACE	CXR+ECG (Combined)		CXR		ECG	
	Mace	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC
Internal	6M	0.571 [0.535, 0.635]	0.608 [0.531, 0.648]	0.631 [0.588, 0.688]	0.512 [0.467, 0.560]	0.568 [0.537, 0.615]	0.600 [0.513, 0.615]
	1yr	0.595 [0.555, 0.638]	0.612 [0.525, 0.675]	0.572 [0.532, 0.613]	0.549 [0.494, 0.592]	0.580 [0.543, 0.633]	0.626 [0.528, 0.640]
	2yr	0.562 [0.511, 0.618]	0.587 [0.512, 0.619]	0.507 [0.459, 0.557]	0.506 [0.457, 0.562]	0.570 [0.534, 0.625]	0.626 [0.534, 0.630]
	5yr	0.596 [0.554, 0.645]	0.617 [0.555, 0.643]	0.509 [0.477, 0.556]	0.510 [0.473, 0.563]	0.579 [0.544, 0.625]	0.628 [0.541, 0.630]
	External (MIMIC)	0.531 [0.368, 0.617]	0.562 [0.349, 0.640]	0.668 [0.571, 0.764]	0.542 [0.430, 0.663]	0.588 [0.444, 0.675]	0.604 [0.422, 0.694]
External (ED)	1yr	0.589 [0.557, 0.649]	0.603 [0.498, 0.666]	0.523 [0.486, 0.548]	0.508 [0.449, 0.589]	0.669 [0.649, 0.690]	0.732 [0.647, 0.695]
Contrastive Alignment Baseline (MedCLIP)							
	Mace	CXR+ECG (Combined)		CXR+ECG (CXR)		CXR+ECG (ECG)	
	Mace	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC
Internal	6M	0.585 [0.583, 0.586]	0.613 [0.610, 0.616]	0.486 [0.484, 0.487]	0.482 [0.480, 0.485]	0.498 [0.497, 0.500]	0.496 [0.494, 0.503]
	1yr	0.593 [0.593, 0.595]	0.628 [0.626, 0.635]	0.494 [0.490, 0.493]	0.492 [0.491, 0.495]	0.500 [0.496, 0.501]	0.499 [0.496, 0.499]
	2yr	0.583 [0.581, 0.585]	0.613 [0.610, 0.613]	0.486 [0.485, 0.488]	0.477 [0.475, 0.479]	0.464 [0.463, 0.467]	0.452 [0.450, 0.454]
	5yr	0.573 [0.570, 0.573]	0.601 [0.597, 0.601]	0.496 [0.495, 0.498]	0.496 [0.495, 0.500]	0.466 [0.463, 0.468]	0.457 [0.456, 0.458]
	External (MIMIC)	0.509 [0.485, 0.511]	0.501 [0.484, 0.519]	0.571 [0.559, 0.584]	0.550 [0.530, 0.560]	0.474 [0.460, 0.484]	0.467 [0.448, 0.484]
External (ED)	1yr	0.522 [0.520, 0.531]	0.488 [0.484, 0.497]	0.517 [0.511, 0.519]	0.528 [0.521, 0.530]	0.489 [0.483, 0.492]	0.487 [0.484, 0.495]
MOSCARD							
With Confounder		CXR+ECG (Combined)		CXR+ECG (CXR)		CXR+ECG (ECG)	
	Mace	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC
Internal	6M	0.681 [0.679, 0.682]	0.733 [0.732, 0.735]	0.670 [0.668, 0.670]	0.716 [0.715, 0.718]	0.673 [0.671, 0.674]	0.723 [0.722, 0.725]
	1yr	0.691 [0.689, 0.692]	0.750 [0.749, 0.752]	0.681 [0.676, 0.683]	0.734 [0.733, 0.736]	0.674 [0.672, 0.676]	0.735 [0.734, 0.738]
	2yr	0.683 [0.681, 0.684]	0.740 [0.738, 0.742]	0.669 [0.665, 0.670]	0.723 [0.723, 0.726]	0.675 [0.673, 0.676]	0.733 [0.731, 0.734]
	5yr	0.678 [0.677, 0.679]	0.735 [0.733, 0.738]	0.669 [0.667, 0.670]	0.717 [0.716, 0.719]	0.675 [0.673, 0.676]	0.730 [0.729, 0.735]
	External (MIMIC)	0.623 [0.614, 0.635]	0.662 [0.649, 0.676]	0.600 [0.590, 0.615]	0.644 [0.624, 0.658]	0.611 [0.594, 0.619]	0.657 [0.633, 0.670]
External (ED)	1yr	0.702 [0.697, 0.704]	0.777 [0.768, 0.783]	0.732 [0.727, 0.734]	0.809 [0.799, 0.811]	0.647 [0.645, 0.653]	0.710 [0.705, 0.714]
Without Confounder		CXR+ECG (Combined)		CXR+ECG (CXR)		CXR+ECG (ECG)	
	Mace	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC	Accuracy AUC
Internal	6M	0.652 [0.649, 0.653]	0.711 [0.707, 0.711]	0.644 [0.643, 0.647]	0.695 [0.692, 0.696]	0.641 [0.640, 0.645]	0.686 [0.684, 0.687]
	1yr	0.662 [0.661, 0.663]	0.724 [0.723, 0.726]	0.656 [0.653, 0.658]	0.713 [0.711, 0.714]	0.647 [0.645, 0.648]	0.694 [0.693, 0.696]
	2yr	0.654 [0.653, 0.656]	0.715 [0.714, 0.718]	0.648 [0.646, 0.649]	0.706 [0.704, 0.708]	0.649 [0.649, 0.652]	0.695 [0.695, 0.698]
	5yr	0.650 [0.647, 0.651]	0.711 [0.708, 0.711]	0.647 [0.645, 0.649]	0.703 [0.701, 0.704]	0.642 [0.640, 0.644]	0.687 [0.685, 0.688]
	External (MIMIC)	0.669 [0.658, 0.677]	0.677 [0.662, 0.690]	0.623 [0.619, 0.642]	0.638 [0.620, 0.649]	0.640 [0.627, 0.650]	0.710 [0.685, 0.713]
External (ED)	1yr	0.737 [0.734, 0.741]	0.810 [0.806, 0.813]	0.772 [0.767, 0.773]	0.837 [0.833, 0.839]	0.673 [0.671, 0.680]	0.753 [0.748, 0.761]

(contrastive) and ALBEF (knowledge distillation) baselines (Table 2). Each multimodal model produces three types of predictions: (1) ‘combined’ is derived from the concatenation of CXR and ECG features after alignment, (2) ‘CXR’ and (3) ‘ECG’ solely on Chest X-ray and ECG features from the aligned space. The best performance on the internal dataset is observed in ‘combined’ predictions from MOSCARD, particularly for MACE prediction at 1 year, with an AUC of 0.750. Interestingly, we found that with MOSCARD, the predictions based solely on ECG features consistently outperform those from the single modality ECG models, highlighting benefits of multimodal integration in improving ECG-based predictions even with single directional attention.

MOSCARD with causality (CXR branch) achieves the best performance on external ED dataset (AUC = 0.837), while MODSCARD (ECG branch) performs best on MIMIC (AUC = 0.71) for MACE prediction at 6 months. These findings highlight how dataset-specific characteristics impact modality effectiveness; however, similar to the internal datasets, multimodal learning enhances performance on both the external ED and MIMIC datasets. Our results indicate that for internal datasets, MOSCARD with the confounder model achieves the best performance against the single modality. Meanwhile, for external datasets, MOSCARD outperforms other models which suggest that de-confounding and causal learning are particularly beneficial for external datasets, likely due to differences in disease prevalence and demographic distributions. Fig. 2 shows that MOSCARD achieved optimal equalized performance across all subgroups including comorbidity and demographics for both internal and external ED data. Deconfounded encoders improved the subgroup performance from the baseline and achieved similar performance on the hold-out but the performance drop on the external ED test. Interestingly, even though CT calcium score (CT imaging marker for coronary atherosclerosis) was not used as a factor for causal intervention, MOSCARD also achieved optimal performance across all the calcium subgroups.

In the context of two-step learning, we applied saliency map interpretation with pixel masking on several interesting cases to explore how MOSCARD reasoning differs from single modality encoding (Fig 2). In both cases, the single modality model mainly relied on external wires and support devices for making predictions. Despite after de-confounding, the focus shifted, but the model’s predictions remained either incorrect or overly confident. In contrast, with ECG guidance, MOSCARD successfully made the correct predictions in both cases, focusing on the heart and lung regions without being distracted by external wires and support devices.

4 Discussion

We introduce the MOSCARD, a novel multimodal causal reasoning model designed to support opportunistic screening of MACE by leveraging existing CXR and 12-lead ECG data, while addressing the challenges posed by causality and confounding factors. Experimental results highlight the effectiveness of MOSCARD

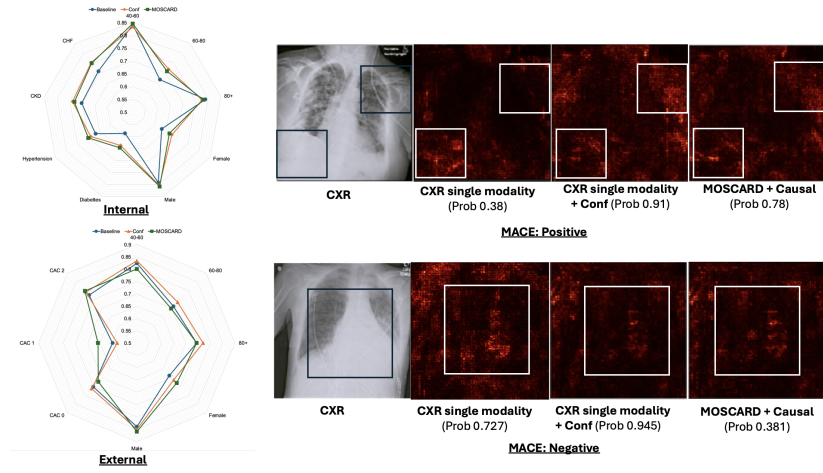


Fig. 2: Spoke plot shows the comparative assessment of MOSCARD (green) against the multimodal co-alignment (blue) and with encoder after de-confounding (orange). Qualitative analysis with saliency maps - (a) Older male patient, true MACE event but single modality failed to predict (prob 0.38), MOSCARD correctly predicted (prob 0.78); (b) Young male patient, no MACE event, single and confounder model predicted false positive but proposed model correctly predicted (prob 0.38).

in providing more accurate and fair predictions for shift in patient populations (high to low risk), compared to traditional single modality and existing multimodal models. De-confounding showed its strength in improving generalizability on external datasets, with the CXR Conf achieving an AUC of 0.778 on the ED dataset and ECG Conf reaching 0.66 AUC on the MIMIC dataset. The MOSCARD with causal reasoning achieved the best performance on external datasets, with the causal version reaching an AUC of 0.837 for MACE at 1 year on the ED dataset. For MIMIC, MOSCARD’s ECG branch performed better (AUC = 0.71). MOSCARD also achieved optimal performance across all subgroups, including those based on comorbidities and demographics, for both internal and external ED datasets.

Rather than relying on patients presenting specifically for cardiovascular evaluations, we propose a novel application for the secondary utilization of multimodal clinical data to enable the early identification of individuals at risk for MACE. This early detection strategy not only facilitates timely interventions and lifestyle modifications but also holds promise in resource-limited settings. On the technical front, our primary contribution is the development of a generalizable multimodal prognostic framework designed to perform robustly across diverse, previously unseen populations with varying prevalence of cardiovascular events. Unlike much of the existing literature, which often overlooks the role of confounders and underlying causal structures in multimodal data, our approach explicitly integrates causal reasoning by incorporating comorbidities and miti-

gating demographic biases using confusion loss. This debiasing strategy enhances the model’s generalizability and effectiveness across heterogeneous populations. The ED dataset includes healthier patients with fewer comorbidities and a lower MACE rate, whereas the MIMIC-IV represents a more complex ICU population with additional comorbidities. After co-alignment, the CXR branch performs best on the ED dataset by identifying subtle signs of early-stage cardiovascular abnormalities, such as early atherosclerosis or pulmonary congestion, while the ECG branch outperforms on the MIMIC-IV patients who are more likely to present with advanced stages of disease based on assessment of electrical activity; however, the co-alignment of CXR and ECG within MOSCARD can learn from both modalities and adjusting for the differences in the populations which help the model to significantly surpassed the performance of single modality. Additionally, our training process is divided into two steps. If one modality is unavailable, the model can still be evaluated using only the first step, which includes confounder debiasing. Due to the limited availability of open-source global datasets with MACE labels, we were unable to assess the model’s generalizability beyond the U.S. healthcare setting, which represents a limitation of our study.

Acknowledgments. The work is supported by NHLBI funded proposal titled *Opportunistic Screening for ASCVD using a Multimodal Deep Learning Risk Prediction Model* grant no. R01 HL167811-02 (PI: Banerjee).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. *Nature Medicine* **28**(9), 1773–1784 (2022)
2. Banerjee, I., Bhattacharjee, K., Burns, J.L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B.N., Shiradkar, R., Gichoya, J.: “shortcuts” causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology* **20**(9), 842–851 (2023). <https://doi.org/https://doi.org/10.1016/j.jacr.2023.06.025>, <https://www.sciencedirect.com/science/article/pii/S1546144023005264>
3. Behar, J.A., Levy, J., Celi, L.A.: Generalization in medical ai: a perspective on developing scalable models (2023), <https://arxiv.org/abs/2311.05418>
4. Brindle, P.M., McConnachie, A., Upton, M.N., Hart, C.L., Smith, G.D., Watt, G.C.: The accuracy of the framingham risk-score in different socioeconomic groups: a prospective study. *British Journal of General Practice* **55**(520), 838–845 (2005), <https://bjgp.org/content/55/520/838>
5. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting shortcut learning for fair medical ai using shortcut testing. *Nature communications* **14**(1), 4314 (2023)
6. Budoff, M.J., Young, R., Burke, G., Carr, J., Detrano, R.C., Folsom, A.R., Kronmal, R., Lima, J.A.C., Liu, K.J., McClelland, R.L., Michos, E.D., Post, W.S.,

- Shea, S., Watson, K.E., Wong, N.D.: Ten-year association of coronary artery calcium with atherosclerotic cardiovascular disease (ascvd) events: the multi-ethnic study of atherosclerosis (mesa). *Eur Heart J* **39**(25), 2401–2408 (2018). <https://doi.org/10.1093/eurheartj/ehy217>
7. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3995–4005 (2021). <https://doi.org/10.1109/ICCV48922.2021.00398>
 8. Correa, R., Pahwa, K., Patel, B., Vachon, C.M., Gichoya, J.W., Banerjee, I.: Efficient adversarial debiasing with concept activation vector — medical image case-studies. *Journal of Biomedical Informatics* **149**, 104548 (2024). <https://doi.org/https://doi.org/10.1016/j.jbi.2023.104548>, <https://www.sciencedirect.com/science/article/pii/S1532046423002691>
 9. Correa, R., Shaan, M., Trivedi, H., Patel, B., Celi, L.A.G., Gichoya, J.W., Banerjee, I.: A systematic review of ‘fair’ ai model development for image classification and prediction. *Journal of Medical and Biological Engineering* **42**(6), 816–827 (2022)
 10. Damen, J.A., Pajouheshnia, R., Heus, P., Moons, K.G., Reitsma, J.B., Scholten, R.J., Hooft, L., Debray, T.P.: Performance of the framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. *BMC medicine* **17**, 1–16 (2019)
 11. of Death Collaborators, G..C.: Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *Lancet* (London, England) **392**(10159), 1736–1788 (2018). [https://doi.org/10.1016/S0140-6736\(18\)32203-7](https://doi.org/10.1016/S0140-6736(18)32203-7)
 12. Dinsdale, N.K., Jenkinson, M., Namburete, A.I.: Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage* **228**, 117689 (2021). <https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117689>, <https://www.sciencedirect.com/science/article/pii/S1053811920311745>
 13. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv. *PhysioNet*. Available online at: <https://physionet.org/content/mimiciv/1.0/> (accessed August 23, 2021) pp. 49–55 (2020)
 14. Lee, N., Bang, Y., Lovenia, H., Cahyawijaya, S., Dai, W., Fung, P.: Survey of social bias in vision-language models. *arXiv preprint arXiv:2309.14381* (2023)
 15. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* **34**, 9694–9705 (2021)
 16. Lloyd-Jones, D.M., Wilson, P.W., Larson, M.G., Beiser, A., Leip, E.P., D’Agostino, R.B., Levy, D.: Framingham risk score and prediction of lifetime risk for coronary heart disease. *The American Journal of Cardiology* **94**(1), 20–24 (2004). <https://doi.org/https://doi.org/10.1016/j.amjcard.2004.03.023>, <https://www.sciencedirect.com/science/article/pii/S0002914904004370>
 17. Pearl, J.: *Causality*. Cambridge university press (2009)
 18. Preiss, D., Kristensen, S.L.: The new pooled cohort equations risk calculator. *Canadian Journal of Cardiology* **31**(5), 613–619 (2015). <https://doi.org/https://doi.org/10.1016/j.cjca.2015.02.001>, <https://www.sciencedirect.com/science/article/pii/S0828282X15000860>
 19. Suh, Y.J., Hong, Y.J., Lee, H.J., Hur, J., Kim, Y.J., Lee, H.S., Hong, S.R., Im, D.J., Kim, Y.J., Park, C.H., Kim, T.H., Yoo, K.J., Choi, B.W.: Prognostic value of syntax score based on coronary computed tomography angiography. *Int J Cardiol* **199**, 460–466 (2015). <https://doi.org/10.1016/j.ijcard.2015.07.067>

20. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
21. Williams, M.C., Moss, A.J., Dweck, M., Adamson, P.D., Alam, S., Hunter, A., Shah, A.S.V., Pawade, T., Weir-McCall, J.R., Roditi, G., van Beek, E.J.R., Newby, D.E., Nicol, E.D.: Coronary artery plaque characteristics associated with adverse outcomes in the scot-heart study. *J Am Coll Cardiol* **73**(3), 291–301 (2019). <https://doi.org/10.1016/j.jacc.2018.10.066>